Spatial Analysis of the Indian Subcontinent: the Complexity Investigated through Neural Networks

Giovanni Fusco, Joan Perez

UMR 7300 ESPACE, CNRS / Université de Nice Sophia Antipolis / Université d'Avignon et des Pays de Vaucluse / Aix-Marseille Université
98 Bd Herriot, BP3209, 06200 Nice (France)

Abstract
India is a very complex space for geographical analysis, above all when the focus of the research is on the rapid transformation of the Indian space, related to urbanization and socioeconomic development. This paper adopts an inductive approach using a database specifically conceived for describing the 640 administrative districts of India between 2001 and 2011. Neural Networks SOM and superSOM approaches are used to cluster districts. Different model options will be presented and a few key points like the importance of prior variable clustering and robust initialization will be highlighted. These key points can be considered as essential prerequisites for any spatial analysis using neural networks. The results of the models show that the Indian space can be meaningfully segmented into a limited number of district profiles, corresponding to particular sub-spaces. Our results show a complex and heterogeneous country, with sub-spaces possessing logics of their own and far away from any cliché.

Keywords: India, Urbanization, Socioeconomic Development, Spatial Clustering, Neural Networks, Self-Organizing Maps.

1. Introduction

1.1 Analyzing Indian Space in the Midst of Socioeconomic Evolution

India is today caricatured as a country with two extremes. On one hand, it is considered as the new Eldorado, the "Shining India", a place where multinationals want to establish themselves due to both substantial increase of consumer market and reduced production costs (Alfaro and Chen, 2009). On the other hand, India is also characterized by overcrowding, major presence of slums and mass poverty, both urban and rural (UN-Habitat, 2001; Dewan Verma, 2002). A dual system could indeed concentrate the growing middle class in selected subspaces connected to the world market, while others would be cut off from significant social and economic development. But, are these extremes truly representative of the diversity of the Indian subcontinent? Increases of standards of living and economic growth are clearly not distributed in a homogeneous way within a territory where the segregation is already worsened by a hermetic caste system. This begs the following questions: how can aggregate measures of socioeconomic development, urbanization and well-being be exploited to grasp, quantify and visualize the complexity of spatial differences within the Indian subcontinent? What are the main drivers affecting these spatial differences?

We thus resorted to AI based algorithms, allowing more freedom in knowledge discovery in databases. A multi-stage Bayesian clustering of Indian districts has already been performed (Perez and Fusco, 2014). The authors of this paper thus employed Self-Organizing Maps (SOMs, Kohonen, 2001) as a good alternative to process a large number of factors while still being able to control the different steps of the analysis. Despites the wide use of Neural Networks in land-use and spatial modeling (Diappi 2004, Roy and Thill 2004, Yan and Thill 2009), little use has been made of these methods to explore a NP-complete problem related to a wide-fast growing country.

In this paper, the different steps of the model will be presented and a few key points like the importance of the factor segmentation and the optimization of cluster initialization will be highlighted in order to understand how results were obtained on spatial clustering and characterization of Indian districts. These key steps can be considered as essential prerequisites for any spatial analysis using neural networks. The results of the model show that the Indian space can be meaningfully segmented into a multitude of district profiles, corresponding to particular sub-spaces. Some of these profiles echo the caricatural opposition between modern emerging

India and poverty stricken marginal backwater regions. But in most cases, our results show a much more complex and heterogeneous country, with sub-spaces possessing logics of their own and far away from any cliché.

The text of the paper is organized as follows. In the next subsection the data and a few working hypotheses underlying our research will be presented. Section 2 presents the Neural Networks methodology used in the research. Section 3 presents the application of this methodology to the clustering of Indian districts. Several clustering models have been used ; their results will be commented both from a statistical and from a geographical point of view. Section 4 will highlight overall conclusion and present perspectives of future research.

1.2 A Database for Inductive Analysis of Indian Space

In order to deal with the complexity of the Indian space, a conceptual model has been developed to inform the selection of 55 spatial indicators (table.1). Once calculated, the indicators make up a geographic database covering aspects of economic activity, urban structure, socio-demographic development, consumption levels, infrastructure endowment and basic geographical positioning within the Indian space. All indicators are calculated at the scale of every district of the Indian Union (640 spatial units in 2011) and on a ten year timeframe (2001-2011), in order to focus on the most recent transformations of the Indian society. An important assumption of the analysis is the pertinence of the district level for the analysis of the Indian subcontinent. With the exception of the largest metropolitan areas (namely Delhi, Mumbai and Calcutta, which are subdivided in several districts), districts are practical observing windows for India's diversity: some are almost completely rural (with practically no urban areas within them), others host several small and mid-sized cities. Another important assumption is the weight of the urbanization patterns within the process of socio-demographic modernization. But without special precautions, comparing the urbanization patterns using raw data from official censuses can lead to misleading results. Official administrative definitions of urban areas do not correspond to consistent geographic content, and the analysis could result in comparing random fragments of urban space. To avoid such statistical bias, the urbanization related indicators of the database had been build using the *e-Geopolis* database (Moriconi-Ebrard, 1994). This research program identifies, localizes and digitizes the built-up areas of the world, using the recommendations published by the United Nations (ESA) for the 1980 census round. In short, 18.366 built up areas were digitized as original polygons in a GIS software. These areas contain 29.209 official

settlements (official census villages and towns of India) have been aggregated at the district level in order to calculate the urban area footprint indicator (Perez, 2013). Several other indicators have been designed specifically for this research like:

- the extended urban areas that take into account the rural space that complements almost-contiguous urban areas and forms a larger settlement structure with them (Perez et al. 2015);
- the distance to tier-1 metropolitan area linking India to the World economy that has been calculated from each district centroid coordinates (Perez et al., 2015) ;
- the residential welfare index of Indian population, corresponding to the percentage of household not suffering from dwelling overcrowding (Perez and Fusco, 2015 **).**

Ta**ble 1** List of the 55 variables used as inputs for clustering of Indian districts.

Variable Name	Unit	Reference Year	Source
Population	Inhabitants	2011	Census of India
Population Evolution (Decadal Growth Rate)*	Percentage points	2001 - 2011	Census of India
Scheduled Caste (SC) Population	Share of Population	2011	Census of India
SC Population Evolution*	Percentage points	2001 - 2011	Census of India
Small Households (HHLDS) (less than 3 peoples)	Share of HHLDS	2011	Census of India
Small HHLDS Evolution	Percentage points	2001 - 2011	Census of India
Big HHLDS (more than 6 peoples)	Share of HHLDS	2011	Census of India
Big HHLDS Evolution	Percentage points	2001 - 2011	Census of India
Children (less than 6 years old)	Share of Population	2011	Census of India
Children Evolution*	Percentage points	2001 - 2011	Census of India
Male ratio	Ratio	2011	Census of India
Male ratio Evolution	Percentage points	2001 - 2011	Census of India
Literacy Rate	Share of Population	2011	Census of India
Literacy Rate Evolution	Percentage points	2001 - 2011	Census of India
Secondary and Tertiary Workers	Share of Workforce	2011	Census of India
Secondary and Tertiary Workers Evolution	Percentage points	2001 - 2011	Census of India
Female w/i Secondary and Tertiary Workers	Share of Sec. and Ter. Workforce	2011	Census of India
Female w/i Tertiary Workers Evolution*	Percentage points	2001 - 2011	Census of India
Motorized Two-wheelers	Share of HHLDS	2011	Census of India
Motorized Two-wheelers Evolution	Percentage points	2001 - 2011	Census of India
Car	Share of HHLDS	2011	Census of India

Car Evolution	Percentage points	2001 - 2011	Census of India
Bicycle	Share of HHLDS	2011	Census of India
Bicycle Evolution	Percentage points	2001 - 2011	Census of India
Phone	Share of HHLDS	2011	Census of India
Phone Evolution	Percentage points	2001 - 2011	Census of India
Bank Account	Share of HHLDS	2011	Census of India
Bank Account Evolution	Percentage points	2001 - 2011	Census of India
None of the following Assets: Car, Phone, TV, Computer, Motorized Two-wheelers.	Share of HHLDS	2011	Census of India
No Assets Evolution*	Percentage points	2001 - 2011	Census of India
Home-Ownership	Share of HHLDS	2011	Census of India
Home-Ownership Evolution	Percentage points	2001 - 2011	Census of India
Home-Ownership for Scheduled castes (SC)*	Share of HHLDS	2011	Census of India
Home-Ownership Evolution (SC)*	Percentage points	2001 - 2011	Census of India
Residential Welfare	Share of HHLDS	2011	Author's work/Census
Residential Welfare Evolution	Percentage points	2001 - 2011	Author's work/Census
Residential Welfare (SC)	Share of SC HHLDS	2011	Author's work/Census
Residential Welfare Evolution (SC)	Percentage points	2001 - 2011	Author's work/Census
Urban Areas Footprint	Share of District surface	2011	e-Geopolis
Number of Urban Areas	Urban Areas	2011	e-Geopolis
Number of Major Urban Areas (< 200k)	Urban Areas	2011	e-Geopolis
Extended Urban Areas Footprint (EUA)	Share of District surface	2011	Author's work/e-Geopolis
Urban Areas within EUA	Share of Urban Area surface	2011	Author's work/e-Geopolis
Size Main EUA	Km^2	2011	Author's work/e-Geopolis
Urban Compactness	Ratio of surfaces UA/EUA	2011	Author's work/e-Geopolis
Administrative Density	Inhabitants / Km^2	2011	Census of India
Urban Area Density	Inhabitants / Km^2	2011	e-Geopolis/Census of India
Distance to Coastline	Km	2011	Author's work
Distance to Rank 1 Metropolitan Area	Km	2011	Author's work
Car manufacturer Point of Sales	Points of sale	2013	Car manufacturer websites
Special Economic Zone*	Hectares	2007	Ministry of Commerce & Industry
Airport Flow	Passengers / Year	2013	Airports Authority of India
Number of Ranked Universities	Universities	2013	Webometrics Ranking of World Universities
Highway distance	Km	2011	OpenStreetMap
Number of Train Stations	Train Stations	2013	OpenStreetMap

* Variables eliminated in model 3.

It appears that the current situation is the result of several centuries of evolution during which India's society has moved from a basically rural civilisation to a (partially) globalized and multi-layered socioeconomic compound. Villages, small and mid-sized cities, metropolitan areas, extended urban areas are the catalysts of India's multi-faceted socioeconomic life.

No indicators were used to trace the belonging of districts to the different states of the Union or to wider cultural or linguistic areas. In this respect, our analysis approach is purely inductive: we want to cluster Indian districts without any prior assumption of wider subspaces within the Indian subcontinent. We thus hope that this dataset used as input will allow us to highlight and identify the main drivers of the socio-demographic modernity through a clustering application using neural Networks.

It should be remarked that 5.8% of the 35200 values of the database were missing for different reasons. Missing values deriving from absence of measurement recordings were inferred through a Bayesian statistical procedure (4.8% of database). The remaining missing values are more a question of non-applicability of indicators (for example welfare indicators for Scheduled Castes in districts having no SC population) and could not be removed. Of the 55 variables describing Indian districts, 37 have normal or almost normal distributions, 1 has a bimodal symmetric distribution and 17 have very asymmetric distributions. Among these, the only variable that clearly shows a power-law distribution (*Airport Flow*) was transformed through a log function for its non-zero values. Subsequently, all variables were scaled and centered through mean and standard deviation, in order to allow variable comparability within the distance function of the SOM algorithms (see further).

2. Methodology

2.1 SOM and superSOM clustering

In the absence of an established theory of spatial disparities in India, classical multivariate factor analysis seemed to us inappropriate in order to explore the complexity of the Indian space. Looking for more freedom in knowledge discovery in databases, we thus resorted to Neural Networks AI based algorithms. The first designed network dates back to the 40's (McCulloch and Pitts 1943) but these methods are widely used only since the 80's. There is a wide range of different kinds of neural networks that can be used for different purposes like prediction, pattern learning, clustering etc. These algorithms possess an astonishing ability to process a large

quantities of data in a quick and efficient way. Nowadays, they have been applied in several areas such as environmental modelling, Image browsing systems, medical diagnosis etc. (Fausett 1994) and, more particularly for us, in urban studies (Diappi 2004).

Self-Organizing Maps (SOM) developed by Teuvo Kohonen (1989, 1999, 2001) are a kind of clustering and pattern recognition Neural Networks that focus on the topological structure of cluster sets by using a neighbourhood function in order to preserve the topological properties of the input space. SOMs analyze input data (where each record corresponds to an input vector) by recursively assigning them to a node of a two-dimensional grid. The n x m grid (the map) has a topological structure: each node has a unique (i,j) coordinate and a certain number of direct neighbours (four or six depending on the geometry of the grid, which can be rectangular or hexagonal). SOM algorithms search the closest map node for each input vector using the square of the minimum Euclidean distance (in heterogeneous databases, variables have to be previously scaled in order to be used by Euclidian distance functions). Map nodes are characterized by a weight vector for the different variables of the analysis. This weight vector evolves during the self-organization process, as input vectors (statistical units) are assigned to the node. Nevertheless, map node weights must be initialized. They can, for example, be set to small standardized values using random initialization (on-line method, Akinduko and Mirkes 2012). Database records are presented to the SOM in random order. The map node with a weight vector closest to a given input vector becomes the best matching unit (BMU) for this record. When the BMU is found, the associated map node gets its weights updated and the input vector under analysis will then be associated with this node. Assigning an input vector to a map node amounts to assigning a record to a cluster. At the difference of K-means, the topological properties of SOMs result in clusters which are organized in terms of reciprocal proximity among them. Indeed, the specificity of the Self-Organizing Map is that when the BMU is found, a radius parameter will allow the update of the neighbouring nodes within this radius. This is particularly useful in order to compare geographic-space proximity and variable-space proximity, as it is often the case in spatial analysis.

The Kohonen package (Wehrens and Buydens, 2007) for R environment (Becker et.al, 1998) implements SOM algorithms. Moreover, this package introduces the Super-Organized Map (superSOM) algorithm that allows using separate layers for different kind of input data. Each layer used in the superSOM algorithm can be seen as a subset of the main database. The aim of these subsets is to gather a predefined number of input vectors together in order to reduce the redundant information.

2.2 Coupling SOM and superSOM clustering

There are several drawbacks in using SOM clustering algorithms directly on database variables. First of all, each input vector possesses the same weight. That is to say, a variable composed of random records will influence the outcomes as much as any other variable. Moreover, the actual ever-increasing data set sizes increases the probability to process redundant information. Lastly, the SOM algorithms can only be performed with one layer of information that does not contain any missing or non-applicable values. This is precisely not the case for our database, where 1% of data are missing. When a single value is missing for a given record, the whole input vector has to be removed. SuperSOM algorithm, on the contrary, can process missing and non-applicable values by removing the records before training the Map. They will be mapped later since they are retained in the data (Wehrens and Buydens, 2007).

The number of layers, their weight as well as the variable grouping in layers within the SuperSOM are usually chosen qualitatively by the operator prior the treatment. These issues can be bypassed by coupling together a SOM and a SuperSOM application. The goal is to previously cluster strongly correlated variables before clustering database records. In our application, we thus first transposed the database matrix, eventually deleting districts (which are now columns) with missing data. The 55 variables used as input vectors are not independent dimensions of the analysis. Performing a standard SOM on these inputs will produce new, not directly observable, synthetic factors, which are linear combinations of the original variables. Subsequently, the original database is transposed again, the rows with missing values are re-entered, and the database is divided into subsets according to the prior standard SOM clustering results. A superSOM clustering of districts can now be carried out, by treating every subset of variables as a distinct layer of spatial information.

2.3 Robust initializations of clustering

Clustering algorithm results often depend on random initializations. As far as SOM and superSOM algorithms are concerned, two random initializations are used: the map node weights (initial values of variables for cluster centers) and the order of database record evaluation for BMU assignment. Wehrens and Buydens (2007) consider that the overall results after several random initializations of the algorithms are remarkably consistent. In our experience, the overall consistency of results of most initializations does not mean that a few initializations could not produce pronounced differ-

ences in clustering results. A problem of robustness of results arises. Random initializations in computer algorithms are always pseudo-random initializations using a particular prime seed. Clustering can thus be carried out iteratively on many seeds. The resulting clusterings can finally be compared and their robustness assessed.

In order to do this, we used the Fowlkes-Mallows similarity index (Fowlkes-Mallows 1983), which is a variant of the well-known Jaccard index. FM similarity index is often used to compare a clustering result to a known "true" clustering, in terms of true positives (TP), false positives (FP) and false negatives (FN), as follows:

$$FM = \sqrt{\frac{TP}{TP+FP} * \frac{TP}{TP+FN}}$$

FM index varies between a minimum of 0 (the given clustering differs in every record assignment from the "true" clustering) and a maximum of 1 (the two clusterings coincide).

In our case, we do not have a "true" clustering as benchmark. Every clustering result, associated with a given random seed, is thus compared to every other one. We can thus calculate a matrix of FM similarity indexes among the clustering results. For every given clustering, we can calculate the average value of the FM indexes in a row, corresponding to the average similarity to all other clustering results. The random seed associated to the clustering having the highest FM average yields the most robust initialization of the SOM / superSOM algorithm.

2.4 An R script to perform combined SOM/superSOM clustering with robust initialization

In order to automate SOM/superSOM clustering, we developed an R script using existing packages (kohonen, class, MASS, dendextend). The script is organized in two parts, as follows. Part one loads the necessary packages, imports and pretreats data and implements automated functions. Data are scaled and centered in order to perform SOM/superSOM clustering on comparable variables. The development of automated functions concerns:

- The generation of a set of prime number for initialization.
- The calculation of the FM-index associated with each prime seed and the selection of the best initialization.
- The evaluation and plotting of a range of indicators in order to validate each step of the model (records optimization, layer optimization, codebook quality etc.).

Part 2 is the script "Main" and includes 3 sections. Its first section performs as many SOM clusterings as prime seeds. The best seed, with the

FM-Similarity index closest to the whole drawing will be automatically selected. This section will be used if a simple SOM clustering of records in a database is sought for.

The second section first performs as many SOM clusterings of variables as prime seeds. The best initialization is selected. The variables are thus grouped in layers according to the clustering output of the best initialization. The layers are weighted according to the number of variables within them and used as inputs for a SuperSOM clustering. A superSOM clustering of records is then produced for every prime seed; the best initialization is retained. To conclude, a one factor ANOVA of the clustering results is performed and non-significant variables are detected.

The third and last section implements the same procedure as above and automatically removes non-significant ANOVA variables in order iteratively perform SOM/superSOM clustering.

3. Application: Clustering Indian Districts

3.1 Presentation of experiments

Within our research, three models of increasing complexity were compared in order to cluster Indian districts. All models were obtained through a seed optimization procedure (robust initialization).

Model 1. The first model uses the 55 variables of the database for direct SOM clustering of districts in a 3x3 hexagonal grid. Analysis of variance is performed at the end of the clustering.

Model 2. The second model performs a two-step analysis. First, the 55 variables are clustered in 16 latent factors through a SOM procedure using a 4x4 hexagonal grid. Later, a superSOM clustering of districts, with a 3x3 hexagonal grid, uses as inputs 16 different data layers, which correspond to the 16 clusters of variables of the SOM clustering. Seed optimization in this model is performed both for variable SOM clustering and for district superSOM clustering. Analysis of variance is performed at the end of district clustering.

Model 3. The third model uses the results of the ANOVA of the second model and eliminates from the database 8 variables for which the F statistics is not significant with error threshold of 0.10 (see table 1). A new SOM clustering of the remaining 47 variables is then carried out with a 4x4 hexagonal grid, followed by superSOM clustering of districts with a 3x3 hexagonal grid.

Figure 1 shows the seed optimization results for the three models. In order to obtain a robust initialization of the SOM and superSOM algorithms, 20 different seeds were tested and the most robust (the ones producing clustering results which are most similar to all the others according to Fowlkes-Mallows similarity index) were selected. Initializations are generally more robust for district clustering (best FM index between 0.89 and 0.93) than for variable clustering (best FM index between 0.67 and 0.69).

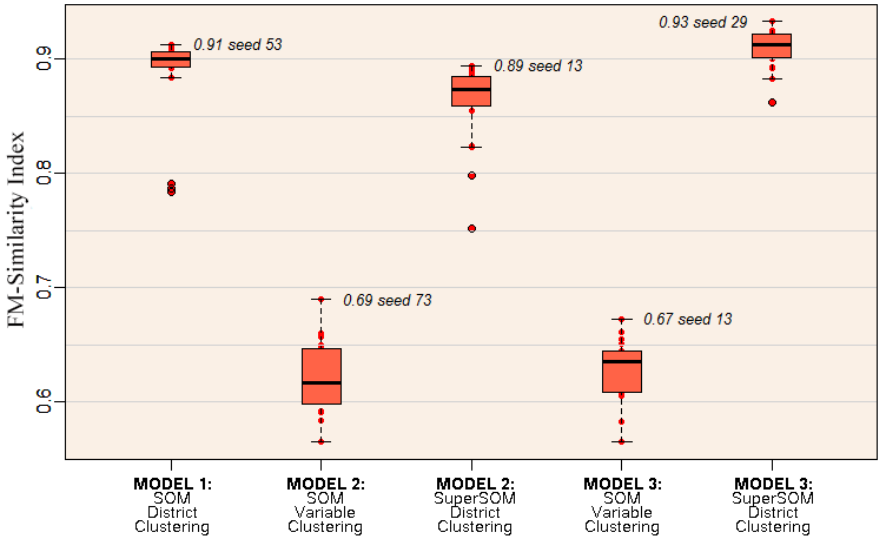

Fig. 1. Boxplots of FM Similarity Index values for 20 model initializations

3.3 Statistical results

The main statistical results of the clusterings from the three models are shown in Figure 2. All three models identify 3 very numerous clusters, each covering more than one hundred Indian districts (min 111 for cluster 3 in model 1, max 181 for cluster 9 in model 2), and 6 more specific ones, with memberships lying between 10 and 70 (min 11 for cluster 8 in model 2, max 67 for cluster 2 in model 3).

The clustering results produced by model 1 seem, from many points of view, of lower quality than those produced by models 2 and 3. Firstly, the SOM algorithm being unable to process missing values, 31 out of 640 districts are deleted and cannot be assigned to any cluster in model 1. Codebook quality is also a way of assessing clustering results. It corresponds to the average distance of records from the cluster center within each cluster. Codebook qualities for all the 9 clusters of districts are particularly medio-

cre in model 1 (the gray color in Figure 2 represents values beyond the 0.05 threshold). Finally, the analysis of variance of clustering results from model 1 identifies 10 non-significant variables (with 0.05 significance threshold).

Models 2 and 3, on the contrary, can cluster all 640 Indian districts and achieve very good codebook qualities for their clusters (all clusters have codebook qualities less than 0.03 and five clusters have values less than 0.02 in both models).

The simple SOM clustering of Indian districts (model 1) is thus clearly outperformed by more refined models, coupling variable SOM clustering and district superSOM clustering. In what follows, we will thus comment only results from models 2 and 3.

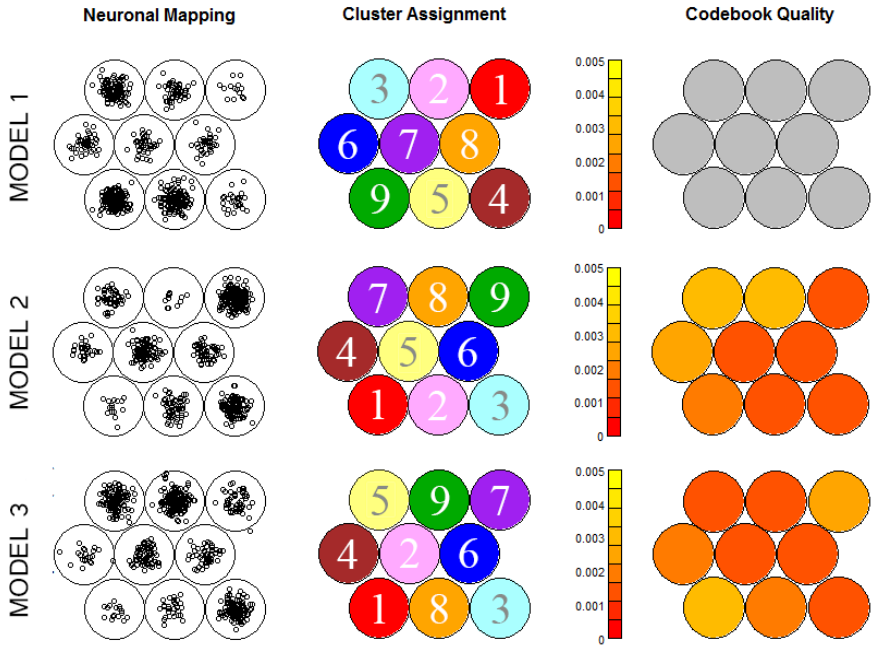

Fig. 2. District Clustering and Codebook Quality for the Three Models

3.4 Geographical results

Models 2 and 3 produce very similar clusterings of Indian districts, with a few exceptions which are worth commenting from a geographic point of view. Equivalences can be found among the two clustering results, which are better highlighted using a common numbering of clusters and repre-

senting cluster profiles for subgroup of clusters (Figure 3). Each subgroup corresponds to a specific geographic reality within the Indian subcontinent and is made of clusters which are (with a few exceptions) topologically close in the 3x3 superSOM grid. Cluster belonging of districts can also be projected in geographic space (Figure 4) and results in a remarkable regionalization of the India Union, which is relatively consistent between the two models.

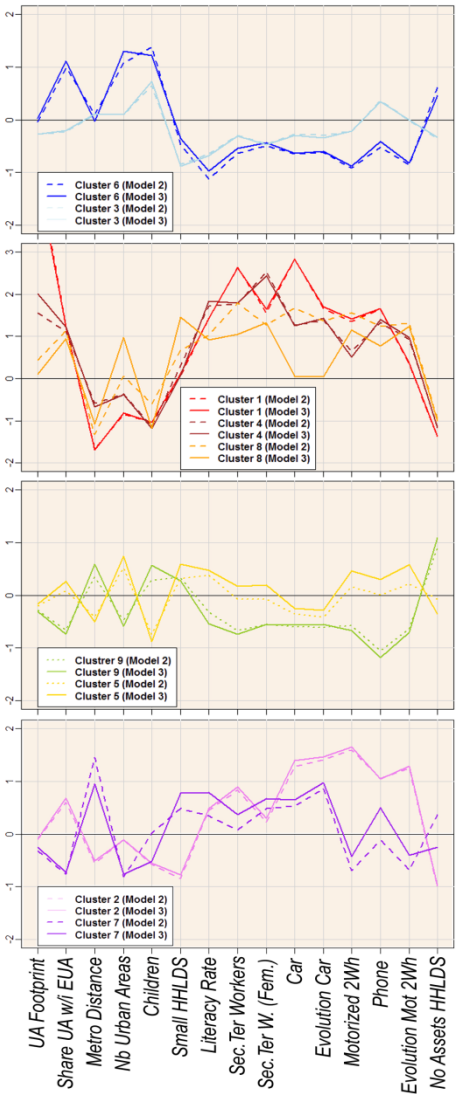

Fig. 3. Cluster Profiles from Model 2 and Model 3 for selected variables.

Poor Traditional Urban India (Clusters 3 and 6).

This grouping encompasses clusters 3 and 6, which are particularly consistent in the two models and always close one to the other in the superSOM grid. Cluster 6 is a cluster of particularly poor and traditional urban India, whose important and dense urbanization falls within larger urban macrostructures. Indeed, most of the districts within this cluster are in the Bihar and West Bengal states, within the large urban macrostructure of the lower Ganges River. Cluster 3 is one of the largest clusters in the Indian subcontinent (117 districts in model 2, 114 in model 3). It is made of traditional urban or urban/rural districts covering most of Uttar Pradesh and Rajasthan, but also smaller areas in Karnataka (in the south) and Jammu and Kashmir (in the north). Socio-demographic tradition (few small households, many big households, important presence of scheduled castes who normally suffer lower living conditions than the general population) and low living standards (even if higher than in cluster 3) are common characteristics of these urban districts which are mainly outside larger urban macrostructures (a few districts are nevertheless part of a large urban macro-structure around Delhi).

Modern Metropolitan India (Clusters 1, 4 and 8).

Clusters 1 and 4 are particularly consistent in the two models. Cluster 1 districts are the forerunners of Indian metropolitan modernity. They are particularly urbanized, often coincide (or are very close to) the most important metropolitan areas, have the most developed economies (with strong presence of secondary and tertiary workers) and very high (by Indian standards) levels of consumption and consumption growth. Not surprisingly the cluster includes all the metropolitan districts of Delhi, the central districts of Mumbai, Kolkata, Hyderabad, Chennai, Chandigarh and Bangalore. These are all big cities particularly well connected to the world economy. Remarkably, model 2 fails to assign Bangalore to cluster 1.

Cluster 4 is a different model of Indian metropolitan development. Districts in cluster 2 are heavily but not completely urbanized, their urban areas are less densely populated, further away from tier-1 metropolitan areas, consumption and consumption growth is high, poverty is low and residential welfare, as well as female presence in services and industry, are even higher than in cluster 1 districts. This metropolitan model is typical of the Kerala state in the south-west, as well as of several regional metropolitan areas in the north-eastern piedmont (like Kampur, Imphal and Bishnupur).
Cluster 8 is slightly different in the two models. In both models, it corresponds to tier-2 metropolitan areas or to dynamic districts which are close

to cluster 1 or cluster 4 districts (metropolitan peripheral districts). The cluster is less numerous in model 2 (only 11 districts), where it corresponds to the closest metropolitan peripheral areas, less urbanized than clusters 1 and 4 but with high consumption levels. In model 3, the cluster also includes many districts of the Tamil-Nadu state in the South. These districts are slightly less affluent and less urbanized than the rest of the cluster but show clear signs of socio-demographic modernity and correspond to urban macrostructures connecting the Kerala conurbation to main metropolitan areas like Chennai and Bangalore.

Non-Urban Well-Off India (Clusters 2 and 7)

Cluster 2 districts are less concerned by urbanization (even if they are not too far from tier-1 metropolitan areas and are often included in larger urban macro-structures) but show high levels of consumption and consumption evolution, low levels of poverty and high presence of secondary and tertiary workers. At the same time, this economic modernity contrasts with socio-demographic traits that are typical of Indian traditional society: low levels of small households, strong presence of big households and high fertility rates, important presence of scheduled castes, low female presence in tertiary and secondary workers, etc. This cluster corresponds to the states of Punjab and Haryana, where we find a model of (relatively) affluent rural or rural/urban India, but also includes districts further south in the Indian subcontinent.

Cluster 7 are mainly rural with important presence of services (mainly in tourism), far away from tier-1 metropolitan areas and disconnected from Indian urban macro-areas. They have nevertheless fairly high levels of consumption and consumption evolution, high residential well-being and typical traits of socio-demographic modernity (small households, low fertility rates, weak or no presence of scheduled castes, even if this is often the result of a different history for north-eastern districts). Geographically, these are districts in northern or north-eastern states close to the Himalaya, or on the islands.

Model 3 also detects a few districts further south (in Karnataka and Maharashtra states), with significant activity in tourism. Less convincingly, model 2 includes in this cluster poorer districts in north-eastern India, increasing the poverty level and reducing the consumption levels of the overall cluster profile.

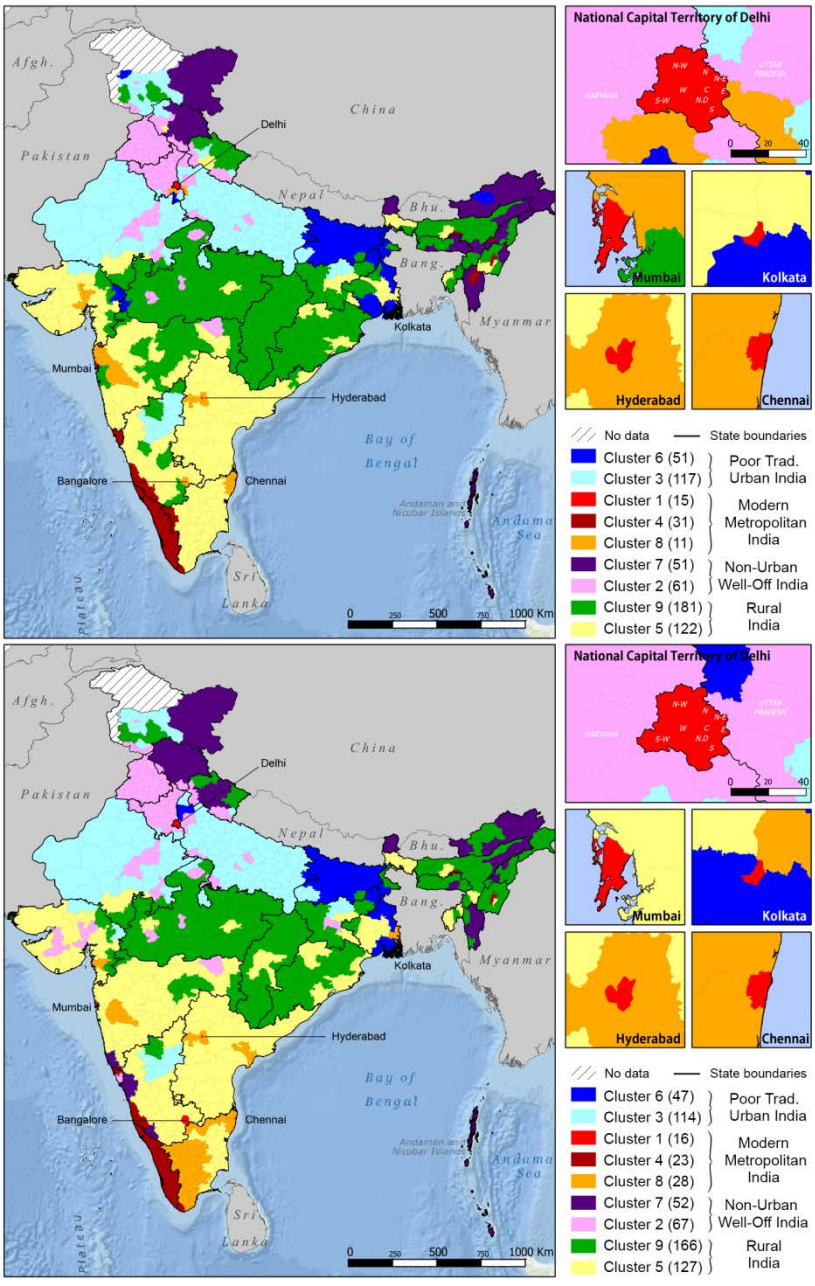

Fig. 4. Clustering of Indian districts in geographic space (top: Model2, Down: Model3)

Rural India (Clusters 5 and 9)

Clusters 5 and 9 are extremely numerous clusters (122-127 and 181-166 districts, respectively), encompassing vast swaths of the Indian subcontinent, mainly rural or weakly urban, for which both superSOM clusterings of models 2 and 3 show little discriminating power. Among the two, cluster 9 is better defined and more consistent in the two models. It corresponds to the poorest districts of rural India, relatively far away from metropolitan areas, with little urbanization, socio-economic backwardness, low consumption levels and widespread poverty. In model 3, it roughly corresponds to the states of Orissa (except for its coastal districts), Chhattisgarh, Jharkhand and Madhya Pradesh. Further north, the poorest districts in Uttaranchal, Jammu and Kashmir and in the seven sister states of India's north-east are also part of cluster 9. Model 2 also integrates several districts in Maharashtra and a few in Karnataka, increasing marginally the average standard of living in the codebook profile. For model 3, the northern border of Maharashtra is a clear-cut geographic limit to poor backward rural India.

Cluster 5 is the "leftover" cluster, where one could say that the superSOM algorithms put districts not belonging to any of the aforementioned clusters. The cluster profile is thus not too far from the average of the 640 districts of the Indian Union (values 0 for every variable). Eventually, these districts are a bit less urbanized, further away from the tier-1 metropolitan areas and with more poverty than the average. For model 3 they are a bit better equipped in phones and motorized-two-wheelers (which are more affordable equipment than cars) and tend to have more small households than the average. Model 2, does not detect these deviances from the Indian average. Geographically, most of the Deccan peninsula is included in this cluster, as well as a few districts in Western Bengal and in the north-east. Indeed, the leftover around-the-average cluster is often a characteristic of any k-means clustering trying to minimize internal variance and maximize inter-cluster variance. Model 3 seems only more able to distinguish between cluster 5 and cluster 9 districts, whereas their two corresponding cluster profiles are closer in model 2, resulting also in a much wider geographic extent for cluster 9.

4. Conclusions

The regionalization of the Indian Union derived from the clustering results of SOM/superSOM coupling is remarkable in its capacity to identify inductively the main spatial structures of the Indian space. It clearly shows

the main spatial oppositions of socioeconomic and urban development within the subcontinent. The lines of divide often (but not always) coincide with state boundaries within the Union. This is the case for states as different as Punjab, Haryana, Uttar-Pradesh, Rajasthan, Madhya Pradesh, Chhattisgarh, Bihar, Kerala, Tamil-Nadu, etc. The very specific cluster of the forerunners of Indian metropolitan modernity is clearly more spatially scattered. SOM/superSOM clustering also identifies other specific models of socioeconomic development in India, more or less linked to urban and metropolitan development. It seems to us that clustering results are less specific for the vast space of intermediate districts, whose situations do not differ constantly and coherently on all variables from the average values of the Indian space.

Validation of clustering results of Indian districts is a difficult task. A few statistical parameters were used for clustering assessment. On a more qualitative basis, clusters have been validated in the paper through expert knowledge of the domain. Field work can also contribute to qualitative clustering validation. Case studies are currently being identified through the clustering results. They will be further investigated in order to enrich the clustering analysis.

Within our models, SOM/superSOM coupling clearly outperforms direct SOM clustering of districts. Removing non-significant variables from the SOM/superSOM analyses also seems to produce slightly better results from a geographic point of view.

Robust initialization of SOM/superSOM clustering is a fundamental step in order to obtain such remarkable results. Other model applications (not shown in this paper) which did not implement robust initialization sometimes produced much less convincing results.

The research presented in this paper can be extended in several directions. The R script can clearly be further developed in order to better automate the analyses. The removal of non-significant variables that we used in order to produce model 3 SOM/superSOM clustering could also be improved. P-values determined through ANOVA do not take into consideration the joint influence of the original database variables on the clusters. Other statistical techniques will thus be explored in order to determine non-significant variables for clustering description.

As far as the analysis of the Indian space is concerned, SOM/superSOM clustering should be carried out using different grid sizes. As with k-means clustering, it will be instructive to compare results when different numbers of clusters are required. It will also be particularly interesting to compare clustering results produces from the same database using different techniques, like Bayesian clustering and decision trees. It is hard to validate the results of a clustering application.

References

Akinduko A.A, Mirkes E.M. (2012) Initialization of Self-Organizing Maps: Principal Components Versus Random Initialization. A Case Study. Machine Learning (stat.ML); Learning (cs.LG) *arXiv*:1210.5873

Alfaro L., Chen M. (2009) *The Global Agglomeration of Multinational Firms*, NBER Working Papers 15576, National Bureau of Economic Research Inc.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole

Dewan Verma G. (2002) *Slumming India: a chronicle of slums and their saviours*, New Delhi: Penguin Books, 183 p.

Diappi L. (Ed.) (2004) *Evolving Cities. Geocomputation in Territorial Planning*. Ashgate, ISBN: 978-0-7546-4194-0, 244 p.

Fausett L. (1994) *Fundamentals of neural networks: Architectures, algorithms, and applications.* New Jersey: Prentice Hall, 469 p.

Fort J - C., Letremy, P. and Cottrell , M. (2002) Advantages and drawbacks of the batch Kohonen algorithm. In M.Verleysen (Ed.) *ESANN'2002 Proceedings, European Symposium on Artificial Neural Networks*, Bruges (Belgium), Bruxelles, Editions D Facto, pp 223 – 230

Fowlkes E. B. , Mallows C. L. (1983) A Method for Comparing Two Hierarchical Clusterings, *Journal of the American Statistical Association*, Vol. 78, No. 383, pp. 553- 569

Kohonen T. (1989) *Self-organizing and associative memory*. (3rd ed.), Berlin: Springer-Verlag.

Kohonen T. (1999) Comparison of SOM Point Densities Based on Different Criteria, *Neural Computation*, 11/1999, p. 2081-2095.

Kohonen T. (2001) *Self-Organizing Maps*, Number 30 in Springer Series in Information Sciences. Springer-Verlag, Berlin, 3rd edition.

Moriconi-Ebrard F. (1994) *Geopolis pour comparer les villes du monde*, Paris, Economica

Perez J. (2013) An Aggregative Approach to Identify and Localize the Hidden Areas of High Consumption in Emerging Countries: India Case Study, Paris, *Conference proceedings*, 29 p.

Perez J, Fusco G. (2014) Inde rurale, Inde urbaine: qualification et quantification de l'aptitude au changement des territoires indiens. In F. Moriconi Ebrard, C. Chatel, J. Bordagi (Eds.) *Fronturb 2014. At the Frontiers of Urban Space - Conference proceedings*, Avignon, ISBN: 978-2-9105-4509-1, halshs-00958799, pp. 316-339

Perez J., Fusco G., Moriconi-Ebrard F.(2015) From Small Towns to Urban Macro-Structure: Quantification and Qualification of Urban and Rural Space in India. *In press*, territoire en mouvement. 29 p.

Perez J., Fusco G. (2015) Residential Welfare at the District Level: Exploring the Standard of Living Gap in India. *Submitted*, Echogeo, 16 p.

Roy J R, Thill J-C. (2003) Spatial interaction modelling, *Papers in Regional Science*, Volume 83, Issue 1, pp 339-361.

United Nations: UN-Habitat. (2001) *Cities in a globalizing world. Global report on human settlements*, edited by United Nations Centre for Human Settlements, 344 p.

Wehrens R, Buydens Lutgarde M. C. (2007) Self- and Super-organizing Maps in R: The kohonen Package, *Journal of Statistical Software*, Vol. 21, Issue 5, Oct 2007.

Yan J, Thill J-C. (2009) Visual data mining in spatial interaction analysis with self-organizing maps, *Environment and Planning B: Planning and Design*, 36(3), pp. 466 – 486.